RNNOTATOR USER'S MANUAL

(Version 2.3)

U.S. Department of Energy
Joint Genome Institute
Lawrence Berkeley National Laboratory

December, 2010

Change Log

Version No.	Date	Revision Description
Rev. 2.3	12/8/2010	Sopra, bambus and velvet added as scaffolding options. Fixed bug during contig splitting when using paired-end, stranded reads. Duplicate read removal and Velvet assembly on by default.
Rev. 2.1	9/9/2010	Support for multiple library types. Oases added as an optional assembler. New artifact filtering options.
Rev. 2.0	8/17/2010	Full support for paired-end reads. Modularization of the program. Integration of accuracy, completeness, contiguity calculation.
Rev. 1.1	4/30/2010	Quality-based read trimming. Better documentation. Duplicate read removal before k-mer filtering.
Rev. 1.0	3/3/2010	Developed the initial version.

TABLE OF CONTENTS

1. Introduction.....1

2. Installation.....1

2.1 Prerequisites.....1

2.2 Installation Details.....2

3. Quick Start.....2

4. User Manual.....3

4.1 Functionality Offered.....3

4.2 Input Formats Accepted.....3

4.3 Options.....4

4.3.1 Read Preprocessing Options.....4

4.3.2 Assembly Options.....5

4.3.3 Assembly Evaluation Options.....5

4.3.4 Advanced Options.....5

5. Galaxy Support.....6

6. License and Citation.....6

7. Contacts.....6

1. Introduction

Comprehensive annotation and quantification of transcriptomes are outstanding problems in functional genomics. Rnnotator is an automated software pipeline that generates transcript models by *de novo* assembly of RNA-Seq data without the need for a reference genome. The contigs produced by Rnnotator are highly accurate and reconstruct full-length genes when transcripts are sequenced sufficiently deep, roughly 30X for a given transcript. Rnnotator was designed to assemble Illumina single or paired-end reads. Rnnotator is also able to incorporate strand-specific RNA-Seq reads into the assembly in order to further improve the assembly.

2. Installation

2.1 Prerequisites

Rnnotator must be run on a 64-bit Linux architecture. Before running Rnnotator the following prerequisites must be installed:

- Blat v. 34 (http://genome.ucsc.edu/FAQ/FAQblat.html#blat3)
- Velvet 1.0.15 (http://www.ebi.ac.uk/~zerbino/velvet/)
- AMOS (http://sourceforge.net/apps/mediawiki/amos/index.php)
- Vmatch 2.0 (http://www.vmatch.de/)
- bwa 0.5.8c (http://bio-bwa.sourceforge.net/)
- MUMmer (http://sourceforge.net/projects/mummer/)
- BioPerl (http://www.bioperl.org)
- Perl modules: Parallel::ForkManager, Tree (http://search.cpan.org/)

Optional prerequisites are:

- Oases 0.1.18 (http://www.ebi.ac.uk/~zerbino/oases/)
- Bambus 2.33 (http://www.cbcb.umd.edu/software/bambus/)
- Sopra 1.0 (dayarian@physics.rutgers.edu) x1 – x4 scripts

2.2 Installation Details

To install Rnnotator, type the following commands into your shell. It will install version 2.3 of the rnnotator package.

1). Gunzip the file and un-tar the file wherever you want the software to reside.

```
$ gzip -cd /path/to/Rnnotator-2.3.tar.gz | tar xf -
```

2). Once the installation is finished, move to the Rnnotator-2.3/ directory.

```
$ cd Rnnotator-2.3
$ perl Makefile.PL PREFIX=/path/to/installation/dir
$ make && make test && make install
```

3). All scripts or programs (64-bit binaries for x86_64 chips for Linux) are in the ./scripts directory. Prior to running Rnnotator, add the scripts directory to the $PATH variable
`e.g., PATH=/path/to/Rnnotator/scripts:$PATH`

When compiling Velvet and/or Oases it is important to set the CATEGORIES and MAXKMERLENGTH directives appropriately for your datasets. The CATEGORIES parameter should be >= the number of libraries types you have. The MAXKMERLENGTH should be proportionally large enough to cover your read length. In other words, if the read length is 75, MAXKMERLENGTH should be at least 69, or so. If you are in doubt, simply set MAXKMERLENGTH to the largest read length when you are compiling Velvet and Oases.

3. Quick Start

- Stranded, paired-end, insert size = 200 bp:

```
$ rnnotator.pl -strP 200 sampleA.fq sampleB.fq
```

- Non-stranded, paired-end, insert size = 200 bp:

```
$ rnnotator.pl -nonP 200 sampleA.fq sampleB.fq
```

- Mixed library types:

```
$ rnnotator.pl –nonS sampleA.fq sampleB.fq –strP 150 sampleC.fq
```

4. User Manual

4.1 Functionality Offered

The Rnnotator pipeline was designed to take advantage of the strengths of existing assemblers, while providing additional functionality to further improve transcriptome assemblies.

Rnnotator takes short read sequences as input and outputs assembled transcript contigs. It consists of three major components: preprocessing of reads, assembly, and post-processing of contigs.

The read preprocessing step may optionally perform several tasks including: removing low-quality reads, low-complexity reads, adapter-containing reads, duplicate reads, reads containing rare k-mers, rRNA containing reads, and read trimming.

After read preprocessing, Rnnotator performs eight assemblies using the assembler of your choice (Velvet, Oases, etc.). Each assembly uses a different hash length for the De Bruijn graph. The assemblies will be run either sequentially or in parallel, depending upon the -n parameter setting. After performing multiple assemblies, Rnnotator removes redundant contigs and further assembles the contigs where significant overlaps are found.

4.2 Input Formats Accepted

Rnnotator accepts FASTQ formatted read files as input. For paired-end reads it is expected that read 1 and read 2 are in the same file and follow one after another in pairs. An example is shown below:

```
@1044:5:1:1071:20262/1
GGTCAATCTCACGATTTGATGGAANAGCTCGCCACCGGGGCAGAGTTCGAGGATGATATAGTAGTATTGACGTGCC
+
bbbbbbbbbbb_a`bbbbbbbbb^BZY[U[bbbbbbbbbab_bab_a`b]b^_bb`]aa`]XT`Z_\__]^_K_BB
@1044:5:1:1071:20262/2
GCAACCAGCGTGCCAACATCCTGAAAGAAGTGCAGATCATGCGCAATCTCGATCACCCCAATATCGTCAAGATGAT
+
bbbabbb_bba^^`bbbbb`bbbc_b`\\b^ab_aaaaa`bcacac_c``aa``ac^`a`aZ^^a^BBBBBBBBBB
```

Any read length produced by Illumina instruments is acceptable for input to Rnnotator. However, 150 bp is the longest read length that has been tested with Rnnotator.

Four library types are supported by Rnnotator: nonS, nonP, strS, and strP, meaning non-strand-specific single-end, non-strand-specific paired-end, strand-specific single-end,

and strand-specific paired-end. At least one library must be given as input to Rnnotator. Multiple FASTQ files can be given for a single library. For example:

```
$ rnnotator.pl -nonS sampleA.fq sampleB.fq sampleC.fq
```

Also, the same library type may be used multiple times. For example, if you have different insert sizes:

```
$ rnnotator.pl -strP 200 sampleA.fq -strP 500 sampleB.fq
```

4.3 Options

4.3.1 Read Preprocessing Options

- `-low_qual on/off`. Remove low quality reads (default: on). Reads are considered low-quality if > 80% of the read length has a quality scores < 20.
- `-low_comp on/off`. Remove low complexity reads (default: on). Low-complexity repeats are defined as homopolymers, di-nucleotide repeats, or tri-nucleotide repeats that compose > 80% of the read length.
- `-adapter on/off`. Remove adapter-containing reads (default: on). Reads are considered adapter reads if the share > 90% identity with the Illumina adapter sequence.
- `-derep on/off`. Remove duplicate reads (default: on). When detecting duplicate reads one mismatch is allowed in the 16 bp hash key. Duplicates are consolidated into consensus reads.
- `-kfilter on/off`. Remove reads containing rare kmers (default: off). Reads are considered to be containing a rare k-mer any k-mer within the read occurs less than `-min_kmer_occur` times using `-kmer_length` as the k-mer length.
- `-trim on/off`. Trim reads to a given length or quality score cutoff (default: off)
- `-trim_len NNN`. Length to trim reads to, when trim is on (default: auto). Auto trimming uses quality scores to determine which length to use when trimming reads. All reads are trimmed to the same length.
- `-rRNA on/off`. Remove rRNA reads (default: off). Removes reads containing ribosomal RNA sequence.

- `-rRNA_fa rDNA.fa`. The ribosomal FASTA file to use when rRNA is on.
- `-rRNA_gs "Genus species"`. If the rRNA FASTA is the silva database then Genus species is used to select sequences from the given Genus and species.

4.3.2 Assembly Options

- `-a assembler`. Assembler to use (velvet, oases) (default: velvet).
- `-s scaffolder` . Scaffolder to use (sopra, bambus, velvet) (default: none).
- `-min_contig_length NNN`. Minimum final contig length (default: 100).
- `-scaffold on/off` . Whether or not to scaffold contigs during velvetg (default: off).

4.3.3 Assembly Evaluation Options

- `-g genome.2bit`. Genome in FASTA or 2bit format, used for reference-based joining and accuracy.
- `-t transcripts.fa`. Transcripts in FASTA format, for checking completeness, contiguity.
- `-ga genes.tab`. Gene annotation in tabular format (name, chrom, strand, start, end, exonStarts, exonEnds). This is used to check % of multigenes and gene fragments in the final contigs.
- `-max_intron NNN`. Maximum intron length, for completeness and contiguity assessment (default: 75000).

4.3.4 Advanced Options

- `-min_kmer_occur NNN`. Minimum number of kmer occurrences for rare kmer filtering (default: 3)
- `-kmer_length NNN`. Kmer length for rare kmer filtering (default: 24).
- `-split_min_cnt NNN`. Minimum depth for transcribed segments when splitting contigs (default: 3)

5. Galaxy Support

Galaxy is a platform for interactive large-scale genome analysis. It is convenient to use the Galaxy platform to create interactive web pages that enable web based analyses instead of using the command line options. Tool configuration and wrapper files which quickly integrate Rnnotator into Galaxy are available upon request.

6. License and Citation

The source code for Rnnotator is available from Lawrence Berkeley National Laboratory under an End-User License Agreement for academic collaborators and under a commercial license for for-profit entities. If you would like to receive this code please contact Virginia de la Puente at vtdelapuente@lbl.gov for details.

If you use Rnnotator, please cite the following paper:

Martin J, Bruno VM, Fang Z, Meng X, Blow M, Zhang T, Sherlock G, Snyder M, Wang Z: **Rnnotator: an automated de novo transcriptome assembly pipeline from stranded RNA-Seq reads**, *BMC Genomics* 2010, **11**:663.

7. Contacts

If you have any questions about Rnnotator, please contact the development team:

- Zhong Wang (ZhongWang@lbl.gov)
- Jeffrey Martin (JAMartin@lbl.gov)
- Xiandong Meng (XiandongMeng@lbl.gov)